LOGISTIC ANALYSIS OF THE IMPACT OF CONTRIBUTING FACTORS ON THE SUCCESS OF STUDENTS OF HIGHER EDUCATION IN QUANTITATIVE COURSES

Vlasta Bahovec
Faculty of Economics and Business, University of Zagreb
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: vbahovec@efzg.hr

Nataša Erjavec
Faculty of Economics and Business, University of Zagreb
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: nerjavec@efzg.hr

Mirjana Čižmešija
Faculty of Economics and Business, University of Zagreb
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia.
E-mail: mcizmesija@efzg.hr

Abstract

The purpose of this paper is to build a logistic regression model where the outcome is the logit of probability of student success in quantitative course and predictor variables are identified factors that contribute to success. The model is based on a random sample of students in professional studies at the University. The applied methodology with obtained results can serve to identify factors that contribute to improving student success in mastering the curriculum, as well as a base for future research on the impact of selected factors on student success in quantitative courses.

Key words: *Logit model, Maximum likelihood estimation, Higher education, Quantitative courses*

1. INTRODUCTION

This paper analyzes the effects of selected factors on the successful mastering of the curriculum of a quantitative course in professional studies at the University. The empirical analysis was conducted over the random sample of students enrolled in 2009/2010 academic year at The Center for Lifelong Learning and Adult Education, who took the exam in Statistics. The goal of the research was to establish whether regular attending to the lectures in Statistics significantly affects the success on exam, and if there are some other variables with significant impact on the success in mastering curriculum. Since the analyzed variable “Success” is a binomial variable, the model of logistic regression is used.

This model, which refers directly to the probability of outcomes, is nonlinear in the parameters. Therefore the logit transformation (log of the odds) is applied to the probability to obtain a linear model. As in a model with a binary dependent variable initial assumptions of a linear regression model are violated and standard least square estimator (LS estimator) produces no meaningful results, logit model is estimated using maximum likelihood estimation method. Once the model is estimated, the respective tests of significance of independent variables are performed. Namely, the significance of individual variables is tested using approximate t-ratios and overall significance of the model is tested using likelihood ratio test. All statistical properties of the logit model are presented together with the regression diagnostic as well. Additionally, some measures of goodness of fit are calculated and compared. Finally, the obtained parameter estimates are interpreted.

The rest of the paper is structured as follows. In section 2 we define the methodology, section 3 presents the main empirical results and finally section 4 concludes.

2. METHODOLOGY

If a dependent variable in the regression model is binomial (or multinomial) the linear regression model is inappropriate for modeling probabilities. For a binary dependent variable with two levels (coded by 0 and 1) and the probabilities $P(Y=1)=p$ and $P(Y=0)=1-p$, the expected value of the dependent variable is $E(Y)=1\cdot p+0\cdot(1-p)=p$. In that case the multiple linear probability model is:

$$p = E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k \tag{1}$$

In model (1), $X' = [X_1, X_2, \ldots, X_k]$ is a vector of k explanatory variables, and $\beta' = (\beta_0, \beta_1, \cdots, \beta_k)$ is a vector of unknown parameters. The linear regression model (1) has certain shortcomings. First of all, although the probabilities take values only in the range $[0;1]$, linear function can take on any value. Apart from that, the observed relationship between the probabilities and explanatory variables is usually nonlinear making the linear regression model inappropriate for the analysis of binary dependent variable. Therefore, a different functional form of the model should be considered.

For the dependent variable Y with two levels the logistic function is:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} = \frac{1}{1+e^{-(\beta_0+X'\beta)}}, \tag{2}$$

where p is the probability that Y=1, β_0 is the intercept and $\beta' = (\beta_1, \cdots, \beta_k)$ is the vector of slope parameters. The equation (2) has the property that the predicted values of the dependent variable are always in the range [0;1] and that the rates of change of the probabilities are not constant for different values of the predictor.

Instead of an equation (2) its logit transformation can be observed. Logit function is the log of the odds (ratio of probability that the event occurs and the probability that it does not occur), *i.e.*

$$\log it(p) = \ln(\frac{p}{1-p}) = \ln(\frac{\frac{1}{1+e^{-(\beta_0+X'\beta)}}}{1-\frac{1}{1+e^{-(\beta_0+X'\beta)}}}) = \ln e^{\beta_0+X'\beta} = \beta_0 + X'\beta \tag{3}$$

In accordance with (3) the logit regression model is:

$$\ln(\frac{p}{1-p}) = \beta_0 + X'\beta + \varepsilon\ . \tag{4}$$

In model (4) p is the probability that the event occurs $p=P(Y=1)$, $\frac{p}{1-p}$ is the odds ratio and $\ln\frac{p}{1-p}$ is the log of odds ratio or logit. $X' = [X_1, X_2, \cdots, X_k]$ is a vector of k explanatory variables, β_0 and $\beta' = (\beta_1, \cdots, \beta_k)$ are unknown parameters. The variable ε is the error term.

The underlying assumptions of the linear regression model are not fulfilled in the logit model. The error term in logit model is not normally distributed and its variance is not homoscedastic. Therefore, the LS estimator is no longer the best linear estimator. The more appropriate estimation method is the maximum likelihood (ML) method, which finds the maximum likelihood parameter estimates by maximizing the likelihood function, which expresses the probability of the observed data as a function of the unknown parameters. The ML procedure is used in an iterative manner, to find the most likely estimates for parameters. ML estimator has the property of consistency, asymptotic normality and asymptotic efficiency. The likelihood value is used for calculating the overall model fit.

When the model is estimated, the convergence of the iterative procedure should be examined first. Then, an overall significance of the model as well as the significance of individual variables has to be tested. Some measures of goodness of fit should be presented and the obtained parameter estimates have to be interpreted. The computer printouts usually contain the following reports: iterations, whole model test, goodness-of-fit statistics, lack of fit test, parameter estimates and effect likelihood ratio (or Wald) tests. For example, the JUMP10/SAS produces the following reports:

Iterations

For the estimated model, with statistically significant independent variables, an iterative estimation process is reported iteration by iteration with evaluated criteria that determine whether the estimated model has converged.

The whole model test

The test shows if the estimated model (the Full model) fits the data better than the model that contains only the intercept (the Reduced model). The whole model test is the likelihood ratio test (*LR*) with the hypothesis:

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_1 : \exists \beta_j \neq 0, \quad j = 1,2,\cdots,k \end{aligned} \tag{5}$$

LR statistic is calculated as minus twice log of ratio of estimated likelihood function ($\hat{L}$) for the model without any predictors and the likelihood function for estimated model. It can be shown (SAS Institute Inc.(2012). *JMP® 10*) that for the large samples under the null hypothesis the test statistic is

$$LR = -2\ln\frac{(\hat{L}(\hat{\beta}_0,0,0,\cdots,0)}{(\hat{L}(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k)} \sim \chi^2(df)\,, \tag{6}$$

where the number of degrees of freedom (*df*) equals the number of parameters[1].

Various goodness-of-fit statistics, such as McFaddens'-R2 (RSquare (U)), corrected Akaike's Information Criterion (AICc) and Bayesian information criterion (BIC) are usually used to test the adequacy of the estimated model. McFaddens'-R2 goodness-of-fit statistic is defined as:

$$RSquare\_(U) = 1 - \frac{\ln \hat{L}(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k)}{\ln \hat{L}(\hat{\beta}_0,0,0,\cdots,0)} \tag{7}$$

It can take values in the range from 0 to 1. Higher values of this statistic are indicative of a good model fit, although they are rare in categorical models. AICc and BIC serve as criteria for model selection among a finite set of models. The smaller values of AICc and BIC the better the model fit.

Lack of Fit Test (Goodness of fit test) gives insight if the variables included in the estimated model give enough information or additional variables have to be added in the model. In other words, it tests if a saturated model is significantly better than the specified model. The test statistic is:

$$LR = -2(\ln \hat{L}_{saturated} - \ln \hat{L}_{Fitted}) \sim \chi^2(df) \tag{8}$$

[1] Continuous independent variables have only one parameter, while models with complex classification effects have one parameter for each anticipated degree of freedom.

The Parameter Estimates with its standard errors and test statistics for the hypothesis that the parameter is zero together with the corresponding p-values and 95% confidence limits are usually presented as well.

The effect likelihood ratio (or Wald) tests test the hypothesis that all the parameters for an individual effect are zero. The significance of a particular parameter can be tested by the Likelihood-ratio Chi-square test or the Wald test. The hypotheses are:

$$H_0 : \beta_j = 0 \qquad (9)$$
$$H_1 : \beta_j \neq 0$$

The Likelihood-ratio Chi-square tests are calculated as twice the difference of the ln($\hat{L}$) of estimated model and the ln($\hat{L}$) of the model without the chosen effect.

$$LR = 2(\ln \hat{L}_{estimated} - \ln \hat{L}_{reduced}) \sim \chi^2(1) \qquad (10)$$

On the other hand, the test statistic for Wald test is:

$$\mathrm{Wald_j} = \left(\frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \right)^2 \sim \chi^2(1) \qquad (11)$$

The Wald statistic is the square of the asymptotic t-ratio, *i.e.* $\hat{t}_j = \hat{\beta}_i / (\hat{\sigma}_j / \sqrt{n})$.

The Effect table

This report shows parameter estimates, standard errors, associated test statistics for the hypothesis that all the parameters for an individual effect are zero and corresponding p-values[2].

Interpretation of the coefficients of the Logit model

An intuitive interpretation of the logit coefficient of the continuous predictor is the effect of the independent variable on the "odds ratio". The odds is the ratio of the conditional probability that Y=1 and the conditional probability that Y=0, for a given $X' = [X_1, X_2, \cdots, X_k]$. From (3) it follows that:

$$Odds(X_1, X_2, \cdots, X_k) = \frac{P(Y=1 | X_1, X_2, \cdots, X_k)}{P(Y=0 | X_1, X_2, \cdots, X_k)} = \frac{\dfrac{1}{1+e^{-(\hat{\beta}_0 + X'\hat{\beta})}}}{1 - \dfrac{1}{1+e^{-(\hat{\beta}_0 + X'\hat{\beta})}}} = e^{\hat{\beta}_0 + X'\hat{\beta}} . \qquad (12)$$

[2] Test statistics are the same as in the Parameter Estimates table if the regressor has only one parameter, but the number of degrees of freedom for ordinal predictors equals *k-1* where k is the number of levels.

Additionally, the odds ratio is the ratio of odds for different values of independent variables. For example, for a one unit increase in the variable X_j, holding the other predictor variables constant at certain value, the odds ratio is:

$$\frac{Odds(X_1, X_2, \cdots, (X_j+1), \cdots, X_k) - Odds(X_1 X_2, \cdots, X_j, \cdots, X_k)}{Odds(X_1 X_2, \cdots, X_j, \cdots, X_k)} = \tag{13}$$
$$= \frac{(1+e^{-(\hat{\beta}_0 + X'\hat{\beta} + \hat{\beta}_j)}) - (1+e^{-(\hat{\beta}_0 + X'\hat{\beta})})}{1+e^{-(\hat{\beta}_0 + X'\hat{\beta})}} = e^{\hat{\beta}_j}$$

Therefore, it can be said that for a one-unit increase in the X_j, holding the other predictor variables constant at a certain value, the expected change in the odds ratio is exponentiated coefficient $\hat{\beta}_j$.

3. EMPIRICAL RESULTS

As an example of the application of the logit model, we try to recognize factors that influence whether a student will successfully pass the exam in Statistics or not. The response variable in the model is Success in Statistics (success=1, failure=0) and the predictor variables of interest are presented in *Table 1.*

Table 1: Predictor variables of interest

Variable name	**Description**	Variable type, measurement unit
Lectures	The presence on lectures in Statistics	Continuous variable
Exercise	The presence on exercises in Statistics	Continuous variable, decile
Math score	Gathered scores in Math	Continuous variable, one point
INF grade	Evaluation on the exam in Informatics	Ordinal variable, values ordered 1-5
Gender	Male, female	Dummy variable, male=1, female=0
Age	Age in years	Continuous variable, one year
Condition	Personal conditions for learning	Ordinal variable, values ordered 1-5
Employed	Whether or not the student is employed	Dummy variable, employed =1, not employed =0

Source: Authors

The model is based on a random sample of 107 students chosen from student population of the Center for Lifelong Learning and Adult Education in Croatia in 2010. Analysis was performed using the statistical software JUMP10/SAS. The estimated model includes only three statistically significant predictors (Exercise, Math score and INF grade), and the model has converged in five iterations. The reported results for estimated model are presented in Table 2. The negative log-likelihood (–LogLikelihood) is the negative sum of natural logs of the observed probabilities. The value of this measure for the full model is 44,450608 and for the model with intercept (without any predictor) it equals 69,611062. The difference between these two values is 25,160455. The likelihood ratio chisquare of 50,32091 with a p-value less than 0.0001 implies that estimated model (as a whole) fits the data significantly better than a model with no predictors.

Table 2: The Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	25,16045482	6	50,32091	4,05E-09
Full	44,45060758			
Reduced	69,6110624			

Source: authors´ calculations

The values of goodness-of-fit statistics are reported in Table 3. The value of RSquare (U) is 0,3614 and indicates that the model accounts for 36,14% of the variation between the two groups of students.

Table 3: Goodness-of-fit statistics

RSquare (U)	0,361443
AICc	104,0325
BIC	121,611
Observations (or Sum Wgts)	107

Source: authors´ calculations

The chi square statistic of 83,35604 with a p-value 0.5907 in the lack of fit test (Table 4) imply that the saturated model is not significantly better than the specified model. On the bases of the obtained result it can be concluded that there is a little to be gained by introducing additional variables in the model.

Table 4: Lack of Fit Test

Source	DF	-LogLikelihood	ChiSquare	Prob>ChiSq
Lack Of Fit	87	41,67802	83,35604	0,590743
Saturated	93	2,772589		
Fitted	6	44,45061		

Source: authors´ calculations

The two Effect table (Table 5), with LR test and Wald test statistics, show that each predictor is significant.

The Table 6 named Parameter Estimates shows parameter estimates, their standard errors, LR test statistics, associated p-values and the 95% confidence interval of the parameters. Both continuous variables exercise and Math score are statistically significant, as is the modalitiy INF_grade[4-3] of ordinal variable INF-grade.

Table 5: Effect Wald Tests and Effect Likelihood Ratio Tests

Effect Wald Tests				
Source	**Nparm**	**DF**	**Wald ChiSquare**	**Prob>ChiSq**
exercise	1	1	9,467796	0,0021*
Math score	1	1	10,79484	0,0010*
INF_grade	4	4	12,88551	0,0118*
Effect Likelihood Ratio Tests				
Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
exercise	1	1	12,12202	0,0005*
Math score	1	1	13,40775	0,0003*
INF_grade	4	4	16,91886	0,0020*

Source: authors´ calculations

Table 6: Parameter Estimates

Term	**Estimate**	**Std Error**	**ChiSquare**	**Prob>ChiSq**
Intercept	-8,491	2,031721	17,46583	2,93E-05
exercise	0,449675	0,146142	9,467796	0,002091
Math score	0,039212	0,011935	10,79484	0,001018
INF_grade[2-1]	3,110337	1,676809	3,440705	0,063609
INF_grade[3-2]	-1,52156	1,317912	1,332925	0,248286
INF_grade[4-3]	2,081381	0,728375	8,165692	0,004269
INF_grade[5-4]	-0,96944	0,851826	1,295224	0,255088

Source: authors´ calculations

The Table 7 named Unit Odds Ratio presents exponentiated coefficients (the odds), exponentiated confidence limits from Parameter Estimates (Table 6) and reciprocal odds for continuous predictors.

Table 7: Unit Odds Ratios

Term	**Odds Ratio**	**Lower 95%**	**Upper 95%**	**Reciprocal**
exercise	1,567803	1,205937	2,166082	0,637835
Math score	1,039991	1,017635	1,067251	0,961547

Source: authors´ calculations

The interpretation of the coefficients of the Logit model for the numeric variable (Table 6) is as follows: The coefficient for **exercise** is $\hat{\beta}_1 = 0{,}4496753$ and exp(0,4496753)= 1,567803. Accordingly, the fitted model shows that, holding **Math scores** constant, the odds of success on the exam in Statistics will increase by 56,78% for the unit increase (one decile or 10%) when attending the exercises in Statistics. This increase does not depend on the value that **exercise** is held at. Similarly, holding variable **exercise** constant**,** the odds of success on the exam in Statistics will increase for about

4% for a one point increase in **Math score**. Namely, since the coefficient for **Math score** is $\hat{\beta}_2 = 0{,}039212$, the exp(0,039212) equals 1,039991.

The 95% confidence limits for **Math score** can be interpreted that with the confidence of 0,95 the success on the exam in Statistics will increase between 1,76% and 6,73% for a unit increase in **Math score,** holding variable **exercise** constant.

The reciprocal odds of 1,039991 is 0,961547, Table 7. That means that for a one point increase in **Math score,** holding Exercise constant, the odds of failure is 0,961547 or that the odds of failure will decrease by 3,84%, since (0,961547-1)·100=-3,8453%.

In addition, the table with odds ratios for ordinal variable INF_grade is presented. Each level of the INF grade variable is compared with other levels. For example, odds ratio for the levels INF_grade 4 and INF_grade 3 equals 8,015535 (Table 8). The result is the same as the result obtained in Table 6. (8,015535 is exponentiated value of 2,0813815 which is the parameter estimate of INF_grade[4-3] in Table 6.)

The coefficients for the ordinal predictor have a slightly different interpretation than the coefficients for the continuous numeric ones. For example, odds ratio for levels 4/3 means that a student with the grade 4 in Informatics has approximately 8,02 times the odds of having the success in Statistic compared to a student with grade 3 in Informatics. Other parameters of the INF grade variable levels are not statistically significant.

Table 8: Part of the table: Odds Ratios for INF_grade

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
4	3	8,015535	0,001735	2,103639	38,22196

Source: authors´ calculations

We also investigated the predicted value of attending the exercises in Statistics for assumed probability of success in Statistics if a student has achieved an average score in Mathematics and passed the exam in Informatics for instance with grade 4. The results for specified probabilities of success in Statistics of 0,6 and 0,7 are presented in Table 9 and Figure 1.

Table 9: Inverse Prediction

Math score 73,4159 INF_grade 4

Specified Probability (success in STAT=1)	Predicted exercise	Lower 95%	Upper 95%
0,6000000	5,220513	0,094784	7,514306
0,7000000	6,203072	2,296340	8,676183

Source: authors´ calculations

The results in Table 9 can be interpreted as follows: For specified probability of 60% of success in Statistics, the estimated model predicts that a student with a mean score in Mathematics (of 73,4159) and an achieved grade 4 in Informatics, must attend 52,21% of exercises in Statistics. If the specified probability of success in Statistics is 70%, the same student have to be present on 62,03% of exercises in Statistics.

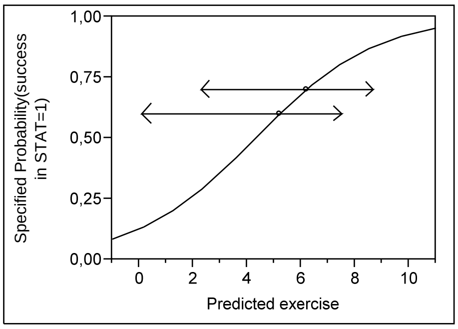

Figure 1: Specified probability

Figure 1 shows Inverse prediction plot. The fitted cumulative probability curve crosses the given 0,6 probability level at the **exercise** value 5,221 which is the inverse prediction. Similarly, the curve crosses the specified 0,7 probability level at the **exercise** value 6,203. Lengths parallel to the horizontal axis, which intersect the curve above the predict values of exercise, are associated 95% confidence intervals.

4. CONCLUSION

The data collected on a random sample of students in the Centre for Lifelong Learning were related to the various socio-demographic variables and scores on the exams on quantitative courses. Using the binomial logistic model the logit of probability of student success in the exam of Statistics in dependence of factors that contribute to success was analyzed. Performed logistic regression over the selected set of variables showed that only the variables exercise, math score and the grade on the exam in informatics are statistically significant predictors. As expected, the regular attending to the lectures in Statistics significantly affects the increase of probability odds of success in mastering course curriculum. In addition, the scores obtained in other quantitative courses significantly increase the probability odds of success in Statistics. The applied methodology with obtained results can serve as a base for the future research on the impact of selected factors on student success in quantitative courses.

REFERENCES

Agresti, A. (1996), *Un Introduction to Categorical Data Analysis*, John Wiley & Sons, New York.

Hair, J.F.Jr., Black, W.C., Babin, B.J., and Anderson, R.E., (2010), *Multivariate Data Analysis,* Pearson Prentice Hall.

http://support.sas.com/rnd/app/stat/papers/exactlogistic2009.pdf, [*Performing Exact Logistic Regression with the SAS System — Revised 2009*].

http://www.ats.ucla.edu/stat/sas/dae/logit.htm, [*SAS Data Analysis Examples Logit Regression*].

http://www.ats.ucla.edu/stat/sas/examples/alr2/default.htm. [*Applied Logistic Regression, 2nd, by Hosmer and Lemeshow*].

SAS Institute Inc. (2004), *SAS/STAT User's Guide*, Version 9.1, Cary, NC: SAS Institute Inc.

Stokes M., Davis C. and Koch G. (2012), *Categorical Data Analysis Using SAS*, Third Edition, SAS Institute Inc, Cary, NC,USA.